

Primitive Generation and Semantic-related Alignment for Universal Zero-Shot Segmentation

Shuting He^{1†} Henghui Ding^{2†✉} Wei Jiang¹
¹Zhejiang University ²Nanyang Technological University
<https://henghuiding.github.io/PADing>

Abstract

We study universal zero-shot segmentation in this work to achieve panoptic, instance, and semantic segmentation for novel categories without any training samples. Such zero-shot segmentation ability relies on inter-class relationships in semantic space to transfer the visual knowledge learned from seen categories to unseen ones. Thus, it is desired to well bridge semantic-visual spaces and apply the semantic relationships to visual feature learning. We introduce a generative model to synthesize features for unseen categories, which links semantic and visual spaces as well as address the issue of lack of unseen training data. Furthermore, to mitigate the domain gap between semantic and visual spaces, firstly, we enhance the vanilla generator with learned primitives, each of which contains fine-grained attributes related to categories, and synthesize unseen features by selectively assembling these primitives. Secondly, we propose to disentangle the visual feature into the semantic-related part and the semantic-unrelated part that contains useful visual classification clues but is less relevant to semantic representation. The inter-class relationships of semantic-related visual features are then required to be aligned with those in semantic space, thereby transferring semantic knowledge to visual feature learning. The proposed approach achieves impressively state-of-the-art performance on zero-shot panoptic segmentation, instance segmentation, and semantic segmentation.

1. Introduction

Image segmentation aims to group pixels with different semantics, e.g., category or instance [11]. Deep learning methods [9, 11, 27, 34, 35, 44] have greatly advanced the performance of image segmentation with the powerful learning ability of CNNs [28] and Transformer [54]. However, since deep learning methods are data-driven, great

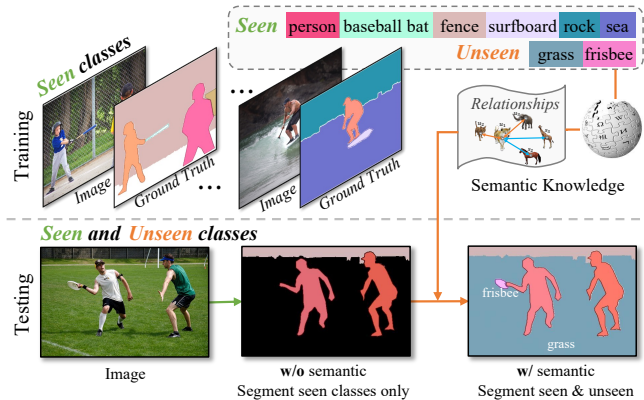


Figure 1. Zero-shot image segmentation aims to transfer the knowledge learned from seen classes to unseen ones (*i.e.*, never shown up in training) with the help of semantic knowledge.

challenges are induced by the intense demand for large-scale labeled training samples, which are labor-intensive and time-consuming. To address this issue, zero-shot learning (ZSL) [36, 47] is proposed to classify novel objects with no training samples. Recently, ZSL is extended into segmentation tasks like zero-shot semantic segmentation (ZSS) [4, 57] and zero-shot instance segmentation (ZSI) [63]. Herein, we further introduce zero-shot panoptic segmentation (ZSP) and aim to build a universal framework for zero-shot panoptic/semantic/instance segmentation with the help of semantic knowledge, as shown in Fig. 1.

Different from image classification, segmentation requires pixel-wise classification and is more challenging in terms of class representation learning. Substantial efforts have been devoted to zero-shot semantic segmentation [4, 57] and can be categorized into projection-based methods [19, 57, 61] and generative model-based methods [4, 25, 38]. The generative model-based methods are usually superior to the projection-based methods because they produce synthetic training features for the unseen group, which contribute to alleviating the crucial bias issue [49] of tending to classify objects into seen classes. Owing to the above merits, we follow the paradigm of generative model-

[†]Equal contribution.

✉Corresponding author (henghui.ding@gmail.com).

based methods to address zero-shot segmentation tasks.

However, the current generative model-based methods are usually in the form of per-pixel-level generation, which is not robust enough in the more complicated scenarios. Recently, several works propose to decouple the segmentation into class-agnostic mask prediction and object-level classification [8, 11, 29, 56]. We follow this strategy and degenerate the pixel-level generation to a more robust object-level generation. What’s more, previous generative works [4, 25, 38] usually learn a direct mapping from semantic embedding to visual features. Such a generator does not consider the visual-semantic gap of feature granularity that images contain much richer information than languages. The direct mapping from coarse to fine-grained information results in low-quality synthetic features. To address this issue, we propose to utilize abundant primitives with very fine-grained semantic attributes to compose visual representations. Different assemblies of these primitives construct different class representations, where the assembly is decided by the relevance between primitives and semantic embeddings. Primitives greatly enhance the expressive diversity and effectiveness of the generator, especially in terms of rich fine-grained attributes, making the synthetic features for different classes more reliable and discriminative.

However, there are only real image features of seen classes to supervise the generator, leaving unseen classes unsupervised. To provide more constraints for the feature generation of unseen classes, we propose to transfer the inter-class relationships in semantic space to visual space. The category relationships obtained by semantic embeddings are employed to constrain the inter-class relationships of visual features. With such constraint, the visual features, especially the synthesized features for unseen classes, are promoted to have a homogeneous inter-class structure as in semantic space. Nevertheless, there is a discrepancy between the visual space and the semantic space [10, 52], so as to their inter-class relationships. Visual features contain richer information and cannot be fully aligned with semantic embeddings. Directly aligning two disjoint relationships inevitably compromises the discriminative of visual features. To address this issue, we propose to disentangle visual features into semantic-related and semantic-unrelated features, where the former is better aligned with the semantic embedding while the latter is noisy to semantic space. We only use semantic-related features for relationship alignment. The proposed relationship alignment and feature disentanglement are mutually beneficial. Feature disentanglement builds semantic-related visual space to facilitate relationship alignment and excludes semantic-unrelated features that are noisy for alignment. Relationship alignment in turn contributes to disentangling semantic-related features by providing semantic clues.

Overall, the main contributions are as follows:

- We study universal zero-shot segmentation and propose **Primitive** generation with collaborative relationship **Alignment** and feature **Disentanglement** learning (**PADing**) as a unified framework for ZSP/ZSI/ZSS.
- We propose a primitive generator that employs lots of learned primitives with fine-grained attributes to synthesize visual features for unseen categories, which helps to address the bias issue and domain gap issue.
- We propose a collaborative relationship alignment and feature disentanglement learning approach to facilitate the generator producing better synthetic features.
- The proposed approach **PADing** achieves new state-of-the-art performance on zero-shot panoptic segmentation (ZSP), zero-shot instance segmentation (ZSI), and zero-shot semantic segmentation (ZSS).

2. Related Work

Zero-shot learning (ZSL) [33, 36, 47, 62] aims to classify images of unseen classes with no training samples via utilizing semantic descriptors as auxiliary information. There are two main paradigms: classifier-based methods that learn a visual-semantic projection [1, 39, 62] and instance-based methods [20, 59] that synthesize fake samples for unseen classes. **Generalized zero-shot learning (GZSL)**, introduced by Scheirer et al. [51], aims to classify samples from both seen and unseen sets. Then, Chao et al. [6] show that the ZSL methods can’t work well in GZSL setting from experiments, due to the feature of overfitting on seen classes. Classification score calibration methods [5, 13, 26, 31] and out-of-distribution detector methods [3, 22] are proposed to alleviate this bias issue.

Image Segmentation is one of the most fundamental computer vision tasks [14, 17, 18, 42, 43]. Deep-learning-based image segmentation methods under a fully supervised manner are extensively studied [9, 11, 15, 16, 27, 34, 35, 44, 56]. However, these methods require a large number of labeled training samples and cannot handle unseen categories that do not appear or are not defined in training data. To address these issues, **Zero-Shot Semantic Segmentation (ZSS)** [4] and **Zero-Shot Instance Segmentation (ZSI)** [63] extend ZSL methods to semantic segmentation and instance segmentation, respectively. In this work, we further introduce **Zero-Shot Panoptic Segmentation (ZSP)** to extend the zero-shot learning to the panoptic segmentation task. There are two main paradigms: projection-based methods [8, 19, 30, 45, 48, 57, 58, 61] and generative-based methods [4, 12, 38]. Projection-based techniques commonly utilize a projection approach to map the visual or semantic features of seen categories onto a shared space. (e.g., visual, semantic, or latent space), and then classify novel objects by measuring the feature

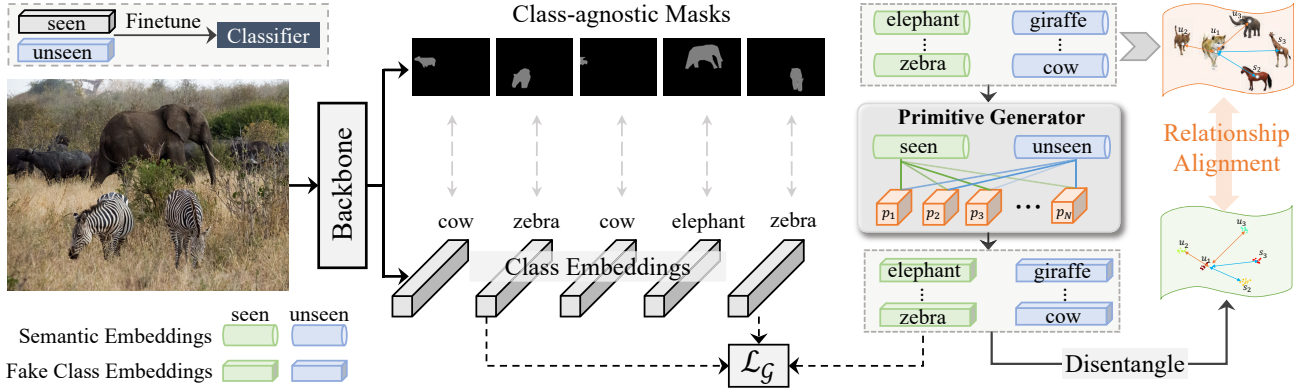


Figure 2. Overview of our approach **PADIing** for universal zero-shot image segmentation. We first obtain class-agnostic masks and their corresponding global representations, named class embeddings, from our backbone. A primitive generator is trained to produce synthetic features (*i.e.*, fake class embeddings). The classifier, which takes class embeddings as input, is trained with both the real class embeddings from image and synthetic class embeddings by the generator. During the training of the generator, the proposed feature disentanglement and relationship alignment are employed to constrain the synthesized features.

similarity in the common space. The generative methods adopt generator to produce synthetic features for unseen classes. However, existing generative works [4, 25, 38] usually learn a direct mapping from semantic embedding to visual features and do not consider the visual-semantic gap of feature granularity. We design a primitive generation and semantic-related alignment approach to universally address zero-shot segmentation, including ZSP, ZSI, and ZSS.

3. Methodology

Fig. 2 illustrates the overview architecture of our proposed approach, **Primitive generation with collaborative relationship Alignment and feature Disentanglement learning (PADIing)**. Our backbone predicts a set of class-agnostic masks and their corresponding class embeddings. Primitive generator is trained to synthesize class embeddings from semantic embeddings. The real & synthetic class embeddings are disentangled to semantic-related and semantic-unrelated features. We conduct the relationship alignment learning on the semantic-related feature. With the synthesized unseen class embeddings, we re-train our classifier with both the real class embedding of seen categories and the synthetic class embedding of unseen categories. The training process is demonstrated in Algorithm 1. The details of each part will be introduced in the following sections.

3.1. Task Formulation

Herein we give the problem formulation of zero-shot image segmentation. There are two spaces, feature space \mathcal{X} and semantic space \mathcal{A} , to represent the visual features of images and semantic representations of categories, denoted as $\mathcal{X} = \{X^s, X^u\}$, $\mathcal{A} = \{A^s, A^u\}$, respectively. The superscript s and u represent the two non-overlapping groups, N^s seen categories and N^u unseen categories,

respectively. We use $\mathcal{Y} = \{Y^s, Y^u\}$ to denote the ground truth label. Y^s is label set of seen group and Y^u is label set of unseen group, $Y^s \cap Y^u = \emptyset$. The training set is constructed from the images that contain any of the N^s seen categories but no unseen categories, which is different from the open-vocabulary paradigm [32, 60]. According to the categories that appear in the testing set, there are two different settings named zero-shot learning (ZSL) and generalized zero-shot learning (GZSL). ZSL only classifies testing samples of unseen categories while GZSL needs to classify testing data of both seen and unseen categories. Zero-shot segmentation is naturally a kind of GZSL since the because images typically contain multiple and diverse categories. In this work, all the zero-shot segmentation tasks are under the GZSL setting unless otherwise specified.

3.2. Primitive Cross-Modal Generation

Due to the lack of unseen samples, the classifier cannot be optimized with features of unseen classes. As a result, the classifier trained on seen classes tends to assign all objects/stuff a label of seen group, which is called bias issue [6]. To address this issue, previous methods [4, 25, 38] propose to utilize a generative model to synthesize fake visual features for unseen classes. However, previous generative zero-shot segmentation works [4, 25, 38] commonly adopt Generative Moment Matching Network (GMMN) [38, 40] or GAN [23], which consist of multiple linear layers as feature generator. Such a generator, though achieves good performance, does not consider the visual-semantic difference of feature granularity. It is well known that image generally contains much richer information than language. Visual information provides very fine-grained attributes of objects while textual information typically provides abstract and high-level attributes. Such difference results in an in-

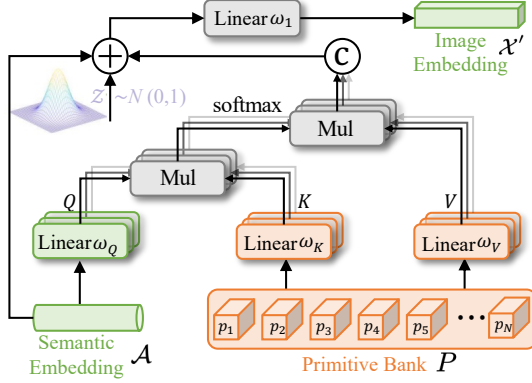


Figure 3. Primitive Cross-Modal Generator. We use lots of learned primitives to represent fine-grained attributes. The generator synthesizes visual features via assembling these primitives according to input semantic embedding.

consistency between visual features and semantic features. To address this challenge, we propose a Primitive Cross-Modal Generator that employs lots of learned attribute primitives to construct visual representations.

As shown in Fig. 3, we build our Primitive Generator with a Transformer architecture. First, a set of learnable primitives are randomly initialized, denoted as $P = \{p_i\}_{i=1}^N$, where $p_i \in \mathbb{R}^{d_k}$ and d_k is the number of channels. These primitives are assumed to contain very fine-grained attributes related to categories, *e.g.*, hair, color, shape, *etc.* Different kinds of assembly of these primitives build different representations for categories. A self-attention is first performed on these primitives to construct relationship graph among these primitives. Next, we utilize two different linear layers ω_K and ω_V to deal with P to obtain the Key and Value for cross attention, denoted as K and V respectively. Then, taking semantic embeddings as Query Q , cross attention is performed as

$$\mathcal{X}' = \omega_1 \left(\text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V + \mathcal{A} + \mathcal{Z} \right), \quad (1)$$

where \mathcal{X}' represents synthetic visual features and \mathcal{Z} denotes random sample with a fixed Gaussian distribution. ω_1 is the linear layer. Different from feature generation via processing semantic embedding with several linear layers, we synthesize visual features via weighted assembling these abundant primitives, which provides much more diverse and richer representations. Moreover, for related categories that share some similarities in semantic space, primitives provide an explicit way to express such similarities. For example, dog and cat both have the attributes of hairy and tail, so the primitives related to hairy and tail show high response to the semantic embedding query of dog and cat. With such primitives that describe fine-grained attributes, we can easily construct different category representations and transfer the knowledge of seen classes to unseen ones.

Algorithm 1 Training of Universal Zero-Shot Segmentation.

- Input:** Image I^s of seen classes, semantic embedding $\mathcal{A} = \{A^s, A^u\}$ from text encoder;
- 1: Pre-train the visual backbone with images of seen classes;
 - 2: Forward image I^s into the trained visual backbone, generating a set of masks M^s and their corresponding class embeddings X^s ;
 - 3: Train primitive generator:
 - a: Forward semantic embedding \mathcal{A} into primitive generator to get synthetic visual feature $\mathcal{X}' = \{X^{s'}, X^{u'}\}$, supervise $X^{s'}$ with X^s by \mathcal{L}_G in Eq. (2);
 - b: Disentangle visual feature $x_i \in \{X^s \text{ or } \mathcal{X}'\}$ into semantic-related feature \hat{x}_i and semantic-unrelated feature \tilde{x}_i ;
 - c: Constrain semantic-related feature \hat{x}_i using alignment \mathcal{L}_A ;
 - 4: Fine-tune the classifier of the visual backbone with X^s and $X^{u'}$.

We follow [40] to define our generator loss \mathcal{L}_G to diminish maximum mean discrepancy between two probability distributions:

$$\mathcal{L}_G = \sum_{f, \hat{f} \in X^s} k(f, \hat{f}) + \sum_{f', \hat{f}' \in X^{s'}} k(f', \hat{f}') - 2 \sum_{f \in X^s} \sum_{f' \in X^{s'}} k(f, f'), \quad (2)$$

where X^s and $X^{s'}$ denote real visual features and synthetic visual features of seen classes, respectively. k is a kernel and $k(f, f') = \exp(-\frac{1}{2\sigma^2} \|f - f'\|^2)$ with bandwidth σ .

When a semantic embedding from unseen group is fed into the trained Primitive Generator, we can get its corresponding synthetic class embedding. We then re-train our classifier with both the real class embedding of seen categories and the synthetic class embedding of unseen categories, which greatly alleviates the bias issue. Besides, such global representations are more robust than per-pixel classification [4, 25, 38, 61, 63] and can thus have a better alignment between visual space and semantic space.

3.3. Semantic-Visual Relationship Alignment

It is well known that relationships among categories are naturally different [7, 38, 55]. For example, there are three objects: apple, orange, and cow. Obviously, the relationship of apple & orange is closer than apple & cow. Class relationships in semantic space are powerful prior knowledge, while the category-specific feature generation does not explicitly leverage such relationships. As shown in Fig. 4, we build such relationships with semantic embeddings and explore to transfer this knowledge to visual space, making semantic-visual alignment in terms of class-wise relationships. By considering the relationship, there are more constraints on the unseen categories' feature generations, to pull or push their distances with seen categories.

Semantic-related Visual Feature However, the visual features are not fully aligned with the semantic representations but contain richer information including semantic-related visual features and also semantic-unrelated visual features. Semantic-unrelated features may have strong

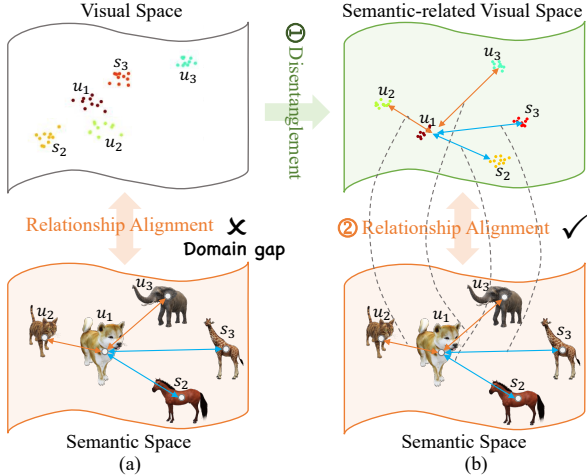


Figure 4. **Relationship alignment.** (a) The conventional relationship alignment. (b) Our proposed two-step relationship alignment. Considering the domain gap, we introduce a semantic-related visual space, where features are disentangled from visual space and have more direct relevance with semantic space. We have the relationship in semantic-related visual space be aligned with semantic space. u_i/s_j refers to unseen/seen category. Taking u_1 dog as an example, we aim to transfer its similarities with {cat, elephant, horse, zebra} from semantic to visual space.

visual clues and contribute to classification, but have low relevance with language semantic representations. Directly aligning semantic embeddings with original visual features would confuse the generator and reduce its generalization to unseen categories. To address this issue, we propose to disentangle the semantic-related visual features and semantic-unrelated visual features. Given a feature x_i , where $x_i \in \mathcal{X}$ is the class embedding from either backbone or our generator, feature disentanglement learns how to disentangle and reconstruct x_i itself. We use encoder $E_{\mathbb{R}}$ to extract semantic-related feature, $\hat{x}_i = E_{\mathbb{R}}(x_i)$. Then, we calculate the correlation score between semantic-related feature \hat{x}_i and semantic embeddings $\mathcal{A} = \{a_1, \dots, a_{N^s + N^u}\}$. $E_{\mathbb{R}}$ is trained with cross-entropy loss as a classification problem to endow semantic-related features \hat{x}_i with discriminative semantic knowledge, *i.e.*,

$$\mathcal{L}_{\mathbb{R}} = - \sum_i \sum_k \mathbb{1}([\hat{x}_i] = k) \log \frac{\exp(\hat{x}_i a_k / \tau)}{\sum_k \exp(\hat{x}_i a_k / \tau)}, \quad (3)$$

where $[\hat{x}_i]$ is the ground truth class index of \hat{x}_i , $\mathbb{1}(\cdot)$ is the indicator function that outputs 1 if the condition is true and 0 otherwise. τ is the temperature parameter.

We use another encoder $E_{\mathbb{U}}$ to extract semantic-unrelated feature, denoted as $\tilde{x}_i = E_{\mathbb{U}}(x_i)$. We suppose the semantic-unrelated features to have the normal distribution $\mathcal{N}(0, 1)$ with zero mean and unit variance [37]. We use KL divergence loss to constrain the distribution range,

$$\mathcal{L}_{\mathbb{U}} = \sum_i D_{KL}[\tilde{x}_i || \mathcal{N}(0, 1)], \quad (4)$$

where $D_{KL}[p||q] = - \int p(z) \log \frac{p(z)}{q(z)}$. Such that each class has its own independent and diverse feature component. To push the network to extract more representative semantic-related features and preserve visual feature information, we reconstruct the feature with a decoder D under ℓ_1 loss:

$$\mathcal{L}_{recon} = \sum_i \|x_i - D(\hat{x}_i, \tilde{x}_i)\|_1. \quad (5)$$

The training objective for feature disentanglement is $\mathcal{L}_{\mathbb{D}} = \mathcal{L}_{\mathbb{R}} + \mathcal{L}_{\mathbb{U}} + \mathcal{L}_{recon}$.

Relationship Alignment Then we conduct relationship alignment between semantic-related visual space and semantic space. We use KL divergence loss to make the similarity of any two semantic-related features \hat{x}_i and \hat{x}_j reach the similarity of their corresponding semantic embeddings $a_{[\hat{x}_i]}$ and $a_{[\hat{x}_j]}$, *i.e.*,

$$\mathcal{L}_{\mathbb{A}} = D_{KL}[\frac{\hat{x}_i \hat{x}_j}{\|\hat{x}_i\| \|\hat{x}_j\|} / \tau || \frac{a_{[\hat{x}_i]} a_{[\hat{x}_j]}}{\|a_{[\hat{x}_i]}\| \|a_{[\hat{x}_j]}\|} / \tau], \quad (6)$$

where $[\hat{x}_i]$ is the ground truth class index of \hat{x}_i , τ is the temperature parameter to control the sharpness of similarity distribution operating on the KL loss. \hat{x}_i^s of the seen group is from either real features or synthetic features while \hat{x}_i^u of the unseen group is from synthetic features by generator only. There are two kinds of alignment, intra-group alignment and inter-group alignment, with different focuses in Eq. (6). When \hat{x}_i and \hat{x}_j are from the same group, *e.g.*, \hat{x}_i^s and \hat{x}_j^s both from seen group, it is intra-group alignment and contributes to extracting better class representations with the relationships as a constraint. When they are from different groups, *e.g.*, \hat{x}_i^s from seen group and \hat{x}_j^u from unseen group, it is inter-group alignment that aims to transfer the relationship knowledge from seen to unseen. Inter-group alignment gives constraints on the relationships of seen and unseen categories, real features and synthetic features. It greatly improves the model's adaptability and generalization to unseen categories.

Collaborative Disentanglement and Alignment Our disentanglement and alignment are complementary and mutually beneficial. On the one hand, disentanglement promotes relationship alignment. With the disentanglement, semantic-related features can be extracted for alignment and semantic-unrelated noises are excluded. On the other hand, relationship alignment facilitates disentanglement. Introducing intra-group and inter-group alignment, class-wise relationship among semantic-related features can be constructed and the discrepancy between semantic-visual feature distributions can be reduced, eventually leading to the improvement of the feature disentanglement.

3.4. Training Objective

Algorithm 1 shows the overall training pipeline of our universal zero-shot segmentation model. First, we pre-train

our segmentation backbone with annotated data from seen classes in a full-supervision manner. Next, We train the primitive generator under the following objective:

$$\mathcal{L}_{total} = \mathcal{L}_G + \lambda(\mathcal{L}_D + \mathcal{L}_A), \quad (7)$$

where λ is the weight to control the importance of the disentanglement and alignment module. Once the generator is trained above, it can generate synthetic features for unseen classes. Together with the real features from seen classes, we can train a new classification layer.

4. Experiments

4.1. Experimental Setup

Implementation Details. The proposed network and all our experiments are implemented based on Pytorch. We utilize CLIP text embeddings [50] and word2vec [46] as our semantic embedding and normalize it with ℓ_2 normalization. CLIP text embeddings are extracted following the previous works [19, 24]. We adopt Mask2Former [11] build upon the ResNet-50 as backbone [28], with 100 queries for both training and inference. Hyper-parameters are consistent with the setting of [11] unless otherwise specified. Encoder E_R and E_U are both multi-layer perceptron (MLP) containing one hidden layer, LeakyReLU activation and dropout. E_D is constructed with two stacked single MLP layers followed by LeakyReLU activation and dropout. We apply SGD optimizer for the parameters of classifier with learning rate 1×10^{-3} , weight decay 5×10^{-4} and momentum 0.9, and Adam optimizer for the parameters of generator, E_R , E_U , and E_D with initial learning rate 2×10^{-4} . The number of the Transformer layers, loss weight λ in Eq. (7), temperature τ , σ are set to 3, 0.002, 0.1, {2, 5, 10, 20, 40, 60}, respectively.

Datasets. We use the popular dataset MSCOCO 2017, which consists training set with 118k images and validation set with 5k images. For panoptic segmentation, 133 classes (80 `thing` classes and 53 `stuff` classes) are included in annotations. For semantic segmentation, COCO-Stuff contains 171 valid classes in total. To get a fair comparison with ZSI [63], we use MSCOCO 2014 for instance segmentation which contains 80k training and 40k validation images.

4.2. Zero-Shot Panoptic Segmentation Task

Because of the high similarities between semantic segmentation and panoptic segmentation, we develop the ZSP datasets by following the previous ZSS works [57]. In order to avoid any information leakage, SPNet selects 15 classes in COCO stuff that do not appear in ImageNet as unseen classes. In COCO panoptic dataset, we find 14 classes overlapped with the 15 ones selected by SPNet and set them as unseen classes, *i.e.*, {cow, giraffe, suitcase, frisbee, skateboard, carrot, scissors, cardboard,

sky-other-merged, grass-merged, playingfield, river, road, tree-merged}, while the remaining 119 classes are set as seen classes. To guarantee no information leakage in the training set, we discard the training images that contain **even one pixel of any unseen classes**. Thus the model is trained by samples of seen classes only with 45617 training images. We use all 5k validation images to evaluate the performance of ZSP. Panoptic and semantic segmentation tasks are evaluated on the union of `thing` and `stuff` classes while instance segmentation is only evaluated on the `thing` classes.

Evaluation Metrics. Under the GZSL setting, the model needs to segment objects/stuff of both seen and unseen classes, which is closer to real-world complicated scenarios. Following previous ZSS [4, 57], ZSD [2], and ZSI [63] tasks, we compute seen metrics, unseen metrics, and the harmonic mean (HM) of seen metrics and unseen metrics as follows,

$$HM = \frac{2 \times P_{seen} \times P_{unseen}}{P_{seen} + P_{unseen}}, \quad (8)$$

where P_{seen} and P_{unseen} denote the seen and unseen metrics, respectively. We use the PQ (panoptic quality) metric [35] which can be viewed as the multiplication of a segmentation quality (SQ) and a recognition quality (RQ). We also report the results on instance segmentation, object detection and semantic segmentation tasks. For instance segmentation and object detection, we use the standard mAP (mean Average Precision) [41] with an IoU threshold of 0.5. For semantic segmentation, we use mIoU (mean Intersection-over-Union) [21].

4.3. Ablation Study

In Tab. 1 and Tab. 2, We perform ablation studies of the proposed **PADing** on MS-COCO dataset under four tasks, including zero-shot panoptic segmentation, zero-shot instance segmentation, zero-shot object detection, and zero-shot semantic segmentation. It is worth noting that the results in Tab. 2 are obtained by the model trained on zero-shot panoptic segmentation task only, which achieves our goal of training a single model for universal zero-shot image segmentation tasks. For simplicity, our ablation analysis mainly focuses on ZSP, because ZSI, ZSD, ZSS have similar trends with ZSP. First, to demonstrate the advantage of introducing generative model, we implement a projection-based segmentation baseline by using CLIP text embeddings as classifier’s weights, similar with ZegFormer-seg [19]. During training, there are 119 text embeddings used in classifier, while during inference, we add another 14 unseen text embeddings into classifier and label each object to one of these 133 classes. As the 2nd row in Tab. 1, there is a strong bias towards seen classes, resulting in extreme low accuracy even zero for unseen group. Next, we construct baseline build upon generative GMMN model following ZS3 [4], which outperforms

Method	G/P	A	D	Seen			Unseen			HM		
				PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ
1) Supervised	X	X	X	43.2	80.8	51.6	0.0	0.0	0.0	-	-	-
2) Projection	X	X	X	43.3	80.9	51.7	0.0	0.0	0.0	0.0	0.0	0.0
3) Baseline	G	X	X	40.1	80.1	48.2	4.9	48.5	5.8	8.7	60.4	10.3
4) P-only	P	X	X	38.9	79.9	46.2	11.5	56.5	13.8	17.7	66.1	21.2
5) P&A	P	✓	X	38.4	79.2	45.8	13.8	52.7	16.4	20.3	63.2	24.1
6) PADing	P	✓	✓	41.5	80.6	49.7	15.3	72.8	18.4	22.3	76.5	26.8

Table 1. Zero-shot panoptic segmentation ablation study results on MSCOCO. G, P, A, D denote GMMN generator, primitive generator, disentanglement, and alignment, respectively.

Method	G/P	A	D	Object Detection (ZSD)			Instance Segmentation (ZSI)			Semantic Segmentation (ZSS)		
				Seen	Unseen	HM	Seen	Unseen	HM	Seen	Unseen	HM
1) Supervised	X	X	X	53.2	0.0	-	53.9	0.0	-	51.2	0.0	-
2) Projection	X	X	X	53.2	12.6	20.4	54.0	12.5	20.3	50.8	1.2	2.3
3) Baseline	G	X	X	52.2	16.5	25.0	52.8	16.2	24.7	50.8	11.6	18.8
4) P-only	P	X	X	52.0	18.5	27.2	52.5	18.4	27.2	50.4	16.7	25.0
5) P&A	P	✓	X	51.9	18.8	27.6	52.3	18.6	27.4	50.2	16.0	24.2
6) PADing	P	✓	✓	52.1	19.6	28.4	52.6	19.2	28.1	50.5	18.5	27.0

Table 2. Ablation study on ZSD, ZSI, and ZSS. G, P, A, D denote GMMN generator, primitive generator, disentanglement, and alignment, respectively. The results validate our goal of training a single model for universal zero-shot image segmentation tasks.

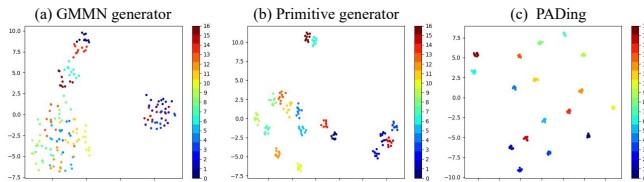


Figure 5. t-SNE visualization of synthesized features for unseen classes. (a) GMMN generator network. (b) The proposed primitive generator. (c) Our **PADing** which further utilizes relationship information based on the primitive generator.

projection-based method by 4.9% in terms of unseen PQ. This phenomenon shows that generative model contributes to solving crucial bias issue.

GMMN vs. Our Primitive Generator. As shown in Tab. 1 and Tab. 2, our primitive generator significantly surpasses GMMN generator by at least 9.0% PQ, 5.7% SQ, and 10.9% RQ for HM metric. This shows that our primitive generator is capable of generating more effective features and the primitives can better grasp the real distribution of visual features compared to the baseline generator GMMN.

#Primitives	100	200	300	400	500	600	700
PQ	7.3	9.1	10.2	11.5	11.5	11.3	11.2

Table 3. Ablation study on the number of primitives.

Number of Primitives. We report the network’s performance with different numbers of primitives in Tab. 3. From the results, increasing the primitive number from 100 to 400 brings a significant performance gain of 4.2%. The performance is a little down when the primitive number is larger than 400, thus we choose 400 as the default setting.

Effectiveness of Alignment. Then, by applying semantic alignment as a constraint to our generator, the HM-PQ is further improved by 2.6%, demonstrating the effectiveness of introducing inter-class relationships inherent from semantic space. Finally, we evaluate the alignment

module with disentanglement, see 6) **PADing** in Tab. 1 and Tab. 2. In comparison to using alignment only, alignment+disentanglement transfers semantic prior knowledge on semantic-related features and consistently brings performance gains of 2.0% HM-PQ, 13.3% HM-SQ, and 2.7% HM-RQ. The significant improvement demonstrates that the semantic-visual discrepancy has been alleviated owing to omitting semantic-unrelated noises. The utilization of disentanglement enables more effective alignment in the separated semantic-related space.

Visualization of synthesized feature representations. To study the properties of our synthesized unseen features and demonstrate the effectiveness of our proposed approach, we employ t-SNE [53] to show the distribution of our synthetic features in Fig. 5. As we can see in Fig. 5 (a), the synthesized features produced by GMMN generator are messy due to the semantic-visual discrepancy. In Fig. 5 (b), when introducing our primitive generator, features belonging to the same class become more compact and features from different classes are highly separable. Furthermore, after applying relationship-alignment constraint on the semantic-related feature, see Fig. 5 (c), features belonging to different classes are farther apart with better-structured distributions, which shows that the structure relationship is embedded into synthetic features and the synthesized unseen features are greatly enhanced with better discrimination.

4.4. Comparison with State-of-the-art ZSS Methods

To further validate the superiority of our approach, we compare it with previous state-of-the-art ZSS methods on the challenging semantic segmentation datasets COCO-Stuff in Tab. 4. It is worth noting that we only report results without self-training and without complicated crop-mask image preprocess utilized for CLIP image encoder for a fair



Figure 6. Qualitative results on the COCO for ZSP. The first row presents input images and the subsequent rows illustrate ground-truth masks, predictions of the baseline, and predictions of our **PAding**.

Method	Embed	Seen IoU	Unseen IoU	HM IoU
SPNet [57]	Word2vec	35.2	8.7	14.0
ZS3 [4]	Word2vec	34.7	9.5	15.0
CaGNet [25]	Word2vec	33.5	12.2	18.2
SIGN [12]	Word2vec	32.3	15.5	20.9
Zsseg-seg [58]	CLIP	38.7	4.9	8.7
ZegFormer-seg [19]	CLIP	37.4	21.4	27.2
PAding(ours)	CLIP	40.4	24.8	30.7

Table 4. Comparison with other ZSS methods on COCO-Stuff.

comparison. We train our model with semantic segmentation annotations. The proposed approach outperforms the previous best method ZegFormer-seg [19] by 3.5% HM-IoU and 3.4% unseen-IoU, demonstrating its effectiveness. It is worth noting that the above methods use ResNet-101 while we only use ResNet-50.

4.5. Comparison with State-of-the-art ZSI Methods

We compare the proposed method with the previous state-of-the-art method ZSI [63] under the Generalized Zero-Shot Instance Segmentation (GZSI) setting in Tab. 5. Our model is trained with instance segmentation annotations for a fair comparison. We achieve new state-of-the-art performance on both 48/17 split and 65/15 split. For example, we surpass ZSI by 7.20% HM-mAP and 5.27% HM-Recall on 48/17 split. It is worth noting that ZSI [63] uses ResNet-101 while we use ResNet-50.

4.6. Qualitative Results

To qualitatively demonstrate the effectiveness of our proposed approach, we visualize some examples of zero-shot panoptic segmentation results in Fig. 6. The second row is ground-truth mask while the third and fourth rows are predicted masks by baseline and our proposed

Split	Method	Seen		Unseen		HM	
		mAP	Recall	mAP	Recall	mAP	Recall
48/17	ZSI [63]	43.0	64.4	3.6	44.9	6.7	52.9
	PAding(ours)	53.0	75.1	8.0	47.5	13.9	58.2
65/15	ZSI [63]	35.7	62.5	10.4	49.9	16.2	55.5
	PAding(ours)	41.8	73.2	13.9	51.3	20.9	60.3

Table 5. Results on GZSI using word2vec embedding.

approach, respectively. We observe that our **PAding** successively finds several unseen classes, *e.g.*, suitcase, grass, frisbee, road, tree, skateboard, that are missed or misclassified by the baseline model. Besides, thanks to the class-agnostic mask generation ability of Mask2Former [11], our results show high-quality masks.

5. Conclusion

We propose primitive generation with collaborative relationship alignment and feature disentanglement learning (**PAding**) as a unified framework to achieve universal zero-shot segmentation. A primitive generator is proposed to synthesize fake training features for unseen classes. A collaborative feature disentanglement and relationship alignment learning strategy is proposed to help the generator produce better fake unseen features, where the former one decouples visual features to semantic-related part and semantic-unrelated part and the later one transfer inter-class knowledge from semantic space to visual space. Extensive experiments on three zero-shot segmentation tasks demonstrate the effectiveness of the proposed approach.

Acknowledgement Shuting He and Wei Jiang were partially supported by National Natural Science Foundation of China (No.62173302).

References

- [1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2
- [2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Eur. Conf. Comput. Vis.*, 2018. 6
- [3] Supritam Bhattacharjee, Devraj Mandal, and Soma Biswas. Autoencoder based novelty detection for generalized zero shot learning. In *ICIP*. IEEE, 2019. 2
- [4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Adv. Neural Inform. Process. Syst.*, 32, 2019. 1, 2, 3, 4, 6, 8
- [5] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Classifier and exemplar synthesis for zero-shot learning. *Int. J. Comput. Vis.*, 128(1), 2020. 2
- [6] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Eur. Conf. Comput. Vis.* Springer, 2016. 2, 3
- [7] Junjie Chen, Li Niu, Liu Liu, and Liqing Zhang. Weak-shot fine-grained classification via similarity transfer. *NeurIPS*, 2021. 4
- [8] Junjie Chen, Li Niu, Siyuan Zhou, Jianlou Si, Chen Qian, and Liqing Zhang. Weak-shot semantic segmentation via dual similarity transfer. *NeurIPS*, 2022. 2
- [9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4), 2017. 1, 2
- [10] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *Int. Conf. Comput. Vis.*, 2021. 2
- [11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1, 2, 6, 8
- [12] Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 9556–9566, October 2021. 2, 8
- [13] Debasmit Das and CS George Lee. Zero-shot image recognition using relational matching, adaptation and calibration. In *IJCNN*. IEEE, 2019. 2
- [14] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Int. Conf. Comput. Vis.*, 2019. 2
- [15] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2
- [16] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2
- [17] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Int. Conf. Comput. Vis.*, 2021. 2
- [18] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 2
- [19] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1, 2, 6, 8
- [20] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 2
- [21] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1), 2015. 6
- [22] Rafael Felix, Ben Harwood, Michele Sasdelli, and Gustavo Carneiro. Generalised zero-shot learning with domain classification in a joint semantic and visual space. In *DICTA*. IEEE, 2019. 2
- [23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11), 2020. 3
- [24] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 6
- [25] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *ACM Int. Conf. Multimedia*, 2020. 1, 2, 3, 4, 8
- [26] Yuchen Guo, Guiguang Ding, Jungong Han, Xiaohan Ding, Sicheng Zhao, Zheng Wang, Chenggang Yan, and Qionghai Dai. Dual-view ranking with hardness assessment for zero-shot learning. In *AAAI*, 2019. 2
- [27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, 2017. 1, 2
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 6
- [29] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2
- [30] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. *Adv. Neural Inform. Process. Syst.*, 33, 2020. 2
- [31] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

- [32] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3
- [33] Pichai Kankuekul, Aram Kawewong, Sirinart Tangruamsub, and Osamu Hasegawa. Online incremental attribute-based zero-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2012. 2
- [34] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2
- [35] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 6
- [36] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2009. 1, 2
- [37] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Eur. Conf. Comput. Vis.*, 2018. 5
- [38] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *Adv. Neural Inform. Process. Syst.*, 33, 2020. 1, 2, 3, 4
- [39] Yan Li, Zhen Jia, Junge Zhang, Kaiqi Huang, and Tieniu Tan. Deep semantic structural constraints for zero-shot learning. In *AAAI*, 2018. 2
- [40] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *ICML*, 2015. 3, 4
- [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 6
- [42] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2
- [43] Chang Liu, Xudong Jiang, and Henghui Ding. Instance-specific feature propagation for referring segmentation. *IEEE Trans. Multimedia*, 2022. 2
- [44] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1, 2
- [45] Fengmao Lv, Haiyang Liu, Yichen Wang, Jiayi Zhao, and Guowu Yang. Learning unbiased zero-shot semantic segmentation networks via transductive transfer. *IEEE Signal Processing Letters*, 27, 2020. 2
- [46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Adv. Neural Inform. Process. Syst.*, 2013. 6
- [47] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. *Adv. Neural Inform. Process. Syst.*, 22, 2009. 1, 2
- [48] Giuseppe Pastore, Fabio Cermelli, Yongqin Xian, Massimiliano Mancini, Zeynep Akata, and Barbara Caputo. A closer look at self-training for zero-label semantic segmentation, 2021. 2
- [49] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, and Xi-Zhao Wang. A review of generalized zero-shot learning methods. *arXiv preprint arXiv:2011.08641*, 2020. 1
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.* PMLR, 2021. 6
- [51] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boulton. Toward open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7), 2012. 2
- [52] Bin Tong, Chao Wang, Martin Klinkigt, Yoshiyuki Kobayashi, and Yuichi Nonaka. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2
- [53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 1
- [55] Maunil R Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *Eur. Conf. Comput. Vis.* Springer, 2020. 4
- [56] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *Eur. Conf. Comput. Vis.* Springer, 2020. 2
- [57] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 6, 8
- [58] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Eur. Conf. Comput. Vis.* Springer, 2022. 2, 8
- [59] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013. 2
- [60] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3
- [61] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 4
- [62] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2
- [63] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 4, 6, 8